NOT TO BE FURTHER REPRODUCED OR COPIED

Splitting up numbers

	Page
Introduction	1-1
1A. Splitting up data	1-3
1B. Two choices of a typical value	1-8
1C. Midmeans	1-11
1D. Clarity of residuals	1-19
1E. Comparing groups	1-22
1F. Some practical considerations	1-26
1G. Back to incomes	1-34
1H. How far have we come?	1-41
*****	
1K. How to put in order	1-44
*****	
1M. Typical values for weighted numbers	1-49
*****	
1P. IEDA processor	1-51
*****	
1R. Arithmetic of splitting	1-52
*****	
1W. Background references	1-54
*****	
1Y. Problems	1-55

1

Splitting Up Numbers

Exploratory data analysis is detective work -- numerical detective work -- or counting detective work -- or graphical detective work.

A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places.

Equally, the analyst of data needs both tools and understanding. It is the purpose of this introduction to provide some of each. Time will keep us from learning about many tools -- we shall try to look at a few of the most general and powerful among the simple ones. We do not guarantee to introduce you to the "best" tools, particularly since we are not sure that there can be unique bests.

Understanding has different limitations. As many detective stories have made clear, one needs quite different sorts of detailed understanding to detect criminals in London's slums, in a remote Welsh village, among Parisian aristocrats, in the cattle-raising west, or in the Australian outback. We do not expect a Scotland Yard officer to do well trailing cattle thieves, or a Texas ranger to be effective in the heart of Birmingham. Equally, very different detailed understandings are needed to be highly effective in dealing with data concerning earthquakes, data concerning techniques of chemical manufacturing,

data concerning the sizes and profits of firms in a service industry, data concerning human hearing, data concerning suicide rates, data concerning population growth, data concerning fossil dinosaurs, data concerning the genetics of fruit flies, or data concerning the latest exploits of molecular biology. A full introduction to data analysis in any one of these fields -- or to any of many others -- would take much more time than our whole introduction.

The Scotland Yard detective, however, would be far from useless in the wild west or the outback. He has certain general understandings of conventional detective work that will help him anywhere. In data analysis there are similar general understandings. We can hope to lead you to a few of them. We shall try.

The processes of criminal justice are clearly divided between the search for the evidence -- in Anglo-Saxon hands the responsibility of the police and other investigative forces -- and the evaluation of the evidence's strength -- a matter for juries and judges. It is helpful to make a similar distinction in data analysis. Exploratory data analysis is detective in character. Confirmatory data analysis is judicial or quasi-judicial in character. Only exploratory data analysis will be our subject here.

Unless the detective finds the clues, judge or jury has nothing to consider. Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider.

Experiments and certain planned inquiries provide some exceptions and partial exceptions to this rule. They do

this because one line of data analysis was planned as part of the experiment or inquiry. Too often to be comfortable, however, restricting one's self to the planned analysis -- failing to accompany it with exploration -- loses sight of the most interesting results.

As all detective stories remind us, many of the circumstances surrounding a crime are accidental or misleading. Equally, many of the indications to be discerned in bodies of data are accidental or misleading. To accept all appearances as conclusive would be destructively foolish, either in crime detection or in data analysis. To fail to collect all appearances because some -- or even most -- are only accidents would, however, be gross misfeasance deserving (and often receiving) appropriate punishment.

Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone -- the first step.

1A. Splitting up data

The aim of exploratory data analysis is to "see into" the data. Sometimes one seeks regularities, sometimes irregularities, but one always tries to penetrate the haze of irrelevance and the fog of variability. One of the greatest single tools in doing this is the separation of the data into parts relevant to separate questions.

Sometimes this separation will require merely throwing pieces of data in the appropriate basket, as when we separate men's weights from women's weights -- or do the same

with strengths of grip, lengths of hair, or college board scores. More often, however, this separation will involve tearing apart individual numbers into two or more parts.

Consider an able bowler, who in the last 10 games scored 193, 214, 185, 198, 203, 207, 191, 184, 210, 192. The two most obvious questions about this bowler's performance are these:

- How well does he generally bowl? What score is typical?
- How consistent is his performance?

Let us, then, take apart these 10 scores in some way (or even in alternative ways) which will tend to separate information about these two questions. We shall not seek "the best way" of taking this data apart. (There may well be no "best way", in any realistic sense.) Instead, we will try to select methods which combine understandability with what experience suggests to be reasonable effectiveness.

What can we do to our bowling scores that will not affect what they tell us about our bowler's irregularity or consistency? Surely subtracting the same constant from all the scores will leave their message about variability undistorted.

Subtraction works well more often than not, when we are trying to make this sort of separation. Other devices sometimes do better. Representing the observation as a percentage of some chosen constant -- which is really the same as looking at the ratio of observation to chosen constant -- is often helpful. We shall see in the next chapter how to turn

this case back to subtraction and thereby gain both simplicity and use of the full kit of tools.

Table 1 gives a collection of what happens to the bowling scores when we subtract constants. In column (2) we have peeled 100 off each score. Clearly this has left the information about variability undisturbed. Equally clearly, we may as well peel off more. Subtracting 150 from each original score, as in column (3) deserves the same comment. After this subtraction what is left ranges from 34 to 64; clearly we would do well to go further.

The next try, in column (4), is the subtraction of 200. What is left tells us about little except variability. If we want to see variability undisturbed by level, 200 is clearly a more purifying choice than 100 or 150. But what of nearby values? Columns (5) and (6) try out 210 and 190 with results that appear somewhat less satisfactory than the subtraction of 200, since the former gives results predominantly made up of negative values, and the latter gives predominantly positive results.

In rather rough terms, we might now answer the two questions as follows:

- The bowler generally rolls about 200; this is a typical score for him.

- His variability is illustrated by column (4) of table 1; he is not infrequently 10 or 15 pins above or below his typical performance, rarely more.

This set of 10 numbers is unlikely to yield much more information than we have just expressed. As was the

table 1 of chapter 1

The 10 Bowling Scores, Raw and With Various Consistent Shifts

(1) Raw scores

(2) Each score less 100

(3) Each score less 150

(4) Each score less 200

(5) Each score less 210

(6) Each score less 190

(7) Each score less 195

(8) Each score less 199

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
193	93	43	-7	-17	3	-2	-6
214	114	64	14	4	24	19	15
185	85	35	-15	-25	-5	-10	-14
198	98	48	-2	-12	8	3	-1
203	103	53	3	-7	13	8	4
207	107	57	7	-3	17	12	8
191	91	41	-9	-19	1	-4	-8
184	84	34	-16	-26	-6	-11	-15
210	110	60	10	0	20	15	11
192	92	42	-8	-18	2	-3	-7

case here, a rather free-and-easy approach to the data often squeezes out most of the available "juice". That this is true is important, for it means that when we need to break new ground in exploring our data -- as we often will -- we can often do quite well with a free-and-easy approach.

This does not mean, however, that we must lose ground if we pick out standard (even hide-bound) processes and plan to use them regularly. Even in exploratory data analysis, and even though their use tends to make us somewhat overrigid, there are often good reasons for choosing and using standard procedures. Two of these reasons stand out:

1) Standard procedures reduce effort and increase the chance that an analysis will be made.

> We had to compute and look at six columns in table 1 in order to decide that 200 was about a reasonable amount to subtract. Both computation and looking involved work that we might better save for other things if we can. A "once through" calculation would be quicker and easier, and is thus more likely to be undertaken.

2) It is much easier to tell people what we have done when this is a standard manipulation.

> Telling people -- partners, supervisors, colleagues, readers, sometimes even one's self -- what was done is always a very important part of data analysis.

1B. Two choices of a typical value

How are we to find a typical value in a usefully standardized way? We are working with numbers. The simplest and easiest things to do with them (or with selected ones of them) are to add them up or to order them. (Counting is even simpler, but unless we order them first it is hard to see how to use counting alone to take us along the way we want to go.)

Those who like to add will have noticed that the sum of the numbers in column (4) is not large -- actually -23. They might well want to pick out a number to subtract so that the sum of what is left -- the sum of the "residuals" -- is zero. To do this, of course, they have only to add up the raw values (finding 1977 in the example) count how many are involved (10 in the example) and divide the sum by the count (finding 1977/10 = 197.7 in the example). The result is often called the arithmetic mean, so that one standard choice is

$$\text{arithmetic mean} = \frac{\text{sum}}{\text{count}}$$

(We could have added small numbers instead of large ones had we taken (-23)/(10) = -2.3 as an adjustment to 200, again finding 197.7.)

Others are naturally more concerned with greater or less, and thus with order in general and with + or - in particular. They will have noticed that column (7) has as many + entries as - ones -- five of each. This seems to them a clear sign that 195 is about right. To compute from the raw

values, they would arrange the scores in order of size, finding this array:

214
210
207
203
198

193
192
191
185
184

They would naturally pick the middle score in such an ordered list. With 10 scores -- or any even number of scores -- there is no single middle number; rather there is a middle pair -- here 193 and 198.

Subtracting any number that is both greater than 193 and less than 198 will leave five residuals + and five - in this example. To be definite we might as well go halfway between and use

$$195.5 = \tfrac{1}{2}(193 + 198)$$

The result is usually called the median of the set of numbers. For it we have:

$$\text{Median} = \begin{cases} \text{middle value, or} \\ \text{halfway between middle pair} \end{cases}$$

Those who take means tend to think poorly of taking medians. At first blush, only one or two values seem to be used. It is perhaps natural -- and certainly incorrect -- for this to seem careless and wasteful. Actually, the median

does quite well in extracting the juice of the matter. It is usually moderately close to the mean in juice-extracting power; sometimes it does a very much better job than the mean. (A professional data analyst is often able to infer -- or judge -- which of the two will do better in dealing with a particular sort of data.)

In a situation like the specific example, where the values to be analyzed seem to behave themselves fairly well, it makes little difference whether we use the mean or the median as the typical value. But suppose our bowler had been "hot" when he rolled the ninth of these games, scoring 300 rather than 210. This change would raise the mean of the ten games by 9, but it would not change the median at all.

Sometimes we do better to regard "wild shots" or outliers -- values such as a 300 would be for this bowler -- as the sort of accidents that should be kept from influencing our typical value -- here our typical score for this bowler. In particular, errors of observation, recording, or copying may be a major cause of such "way out" values. If the ninth string had come to us as 400 or 310 we would have known that an error was involved. (The maximum possible score in bowling is 300.) When it is 300 we can wonder about an error.

When we feel that "far out" values are likely, the median is almost always a better choice than the mean. Even when we only fear wild values that are far out but not way out -- unusual but not errors -- the median is likely to be a better choice.

Sometimes, especially when a typical value is really a step to a total -- for instance when one calculates an average loss for a certain kind of insurance policy -- it is important to let such wild shots have as much effect as any other observation. Many golf championships depend on a total count of strokes; a pitcher's earned run average includes his catastrophic failures. When this is the case, using the mean is to be preferred.

1C. Midmeans

We wrote earlier about the advantages of standardization. Using "the median or the mean" is a long step toward standardization. But having even two alternatives is incomplete standardization. Here, as in so many places in data analysis, complete standardization would be a mistake. There will be places where we need a mean, others where we need a median. We cannot hope for complete standardization.

We can, however, seek out a single compromise for use in a wide variety of circumstances. There are many rather plausible candidates -- we shall select what is perhaps the simplest of all such.

In the ordered list of 10 scores, those most subject to question or laying aside are the extreme scores. The median comes about, in one interpretation, by laying aside some of the highest values and some of the lowest values, almost $\frac{1}{2}$ of all values in each case. A reasonable compromise between doing this and laying aside nothing, as the mean does, is to lay aside about $\frac{1}{4}$ of all the values at each end. The easiest thing to do with what then is left is to take a mean.

Table 2 shows the computation of the easy midmean, where

easy midmean = mean of remainder (after about $\frac{1}{4}$ removed from each end)

for our 10 bowling scores. (Here, as the alternatives show, we have taken "about $\frac{1}{4}$ of 10" namely "about 2.5" as "probably 3, maybe 2").

When one is:

- dealing with a reasonable number of observations
- making an analysis for one's own information
- using paper and pencil

the easy midmean, just described, is likely to be a good choice. When there are few observations, or the results are to be shown to strangers, or an automatic computer is to do all the work, it is usually wise to calculate the exact midmean, defined as

exact midmean = mean of remainder (after exactly $\frac{1}{4}$ has been removed from each end)

In our example with 10 observations, this means removing 2.5 observations from each end. How do we do this?

Removing two observations at each end would give us

$$\frac{191 + 192 + 193 + 198 + 203 + 207}{1 + 1 + 1 + 1 + 1 + 1} = 197.3$$

Removing three observations would give us

$$\frac{192 + 193 + 198 + 203}{1 + 1 + 1 + 1} = 196.5$$

table 2 of chapter 1

Calculation of Easy Midmean

(1) Values as observed.

(2) Values ordered.

(3) Values after setting aside about $\frac{1}{4}$ from each end.*

(4) Calculation of mean of remaining values (setting aside parenthesized values).

(1)	(2)	(3)	(4)
193	214		
214	210		
185	207	(207)	
198	203	203	
203	198	198	
207	193	193	Midsum = 786
191	192	192	Midcount = 4
184	191	(191)	Midmean = $\frac{786}{4}$ = 196.5
210	185		
192	184		

* Values in ( ) should be either both set aside (as done) or both kept (a possible alternative).

The natural halfway point is

$$\frac{\frac{1}{2}(191) + 192 + 193 + 198 + 203 + \frac{1}{2}(207)}{\frac{1}{2} + 1 + 1 + 1 + 1 + \frac{1}{2}} = 197.0$$

and corresponds to giving "half weight" to those observations that are "half removed". We can give "1/4 weight" or "3/4 weight" similarly.

In those circumstances where a single standard should or must be used -- almost come what may -- a midmean is usually a good choice -- a plausible compromise between mean and median. Where the extra effort its computation requires is not important, the exact midmean usually pays its way -- except sometimes for batches of size 3, where the eay midmean deals much more effectively with "way out" values.

Names and patterns

We are going to come back, again and again, to the end quarters and the middle half of a set of numbers. We need both short, simple words and simple patterns of layout that are easy to grasp and easy to remember. There have been no standard English terms, though Germans sometimes refer to "die fliegende Viertel" -- the flying quarters. We shall borrow from this, and call the end quarters the "wings" and the middle half the "body". (What are called tails are usually only a small part of the wings.)

Wings are naturally joined to the body by "hinges". If coming in $\frac{1}{4}$ of the batch size from the end leaves us in the middle of an observation -- if two observations have fractional weight in the exact midmean -- we call these divided

observations the hinges. If coming in $\frac{1}{4}$ of the batch size covers only whole observations, we call the midpoints between the innermost wing observations and the outermost body observations the hinges.

Having the words, what of the layout? Table 3 suggests a layout, illustrating it for:

- all 10 bowling scores.
- the first 8 in order of time.

If we like to use this layout -- or some similar one -- we can make computation of either easy or exact midmeans quite easy.

Skeleton midmeans

When we have many numbers to combine by hand, finding a midmean, easy or exact, can cause us to add up long columns of figures. It often pays to avoid this by calculating the skeleton midmean for which we add up only four values:

- the lower hinge
- the median
- the median again
- the upper hinge

dividing the sum by 4. When there are an even number of observations in the sample, twice the median is just the sum of the two center values, so we can add the two in, never bothering to form the median. For the 10 bowling scores, this gives

$$\text{skeleton midmean} = \frac{191 + 193 + 198 + 207}{4} = 197.2$$

table 3 of chapter 1

Wings, Hinges, Body, and Midmeans for the Bowling Scores: All 10 and the First 8

	Wings	Hinges	Body	Midmeans
(a) ALL 10 BOWLING SCORES:				
	214 210			Easy midmean $= \frac{786}{4} = 196.5$
		207 (keep $\frac{1}{2}$)	203 198 193 192	Exact midmean* $= \frac{\frac{1}{2}(398) + 786}{5}$ $= \frac{985}{5} = 197.0$
	185 184	191		(Mean = $\frac{793+398+786}{10} = 197.7$)
Sums:	793	398	786	
Counts:	4	2	4	
(b) THE FIRST 8 BOWLING SCORES (210 AND 192 OMITTED)				
	214 207	(205)		Easy midmean $= \frac{785}{4} = 196.25$
		(keep)	203 198	
		(none)	193 191	Exact midmean $= \frac{785}{4} = 196.25$
	185 184	(188)		(Mean = $\frac{790+785}{8} = 196.9$)
Sums:	790	**	785	
Counts:	4	**	4	

*Coefficient $\frac{1}{2}$ because $10/4 = 2\frac{1}{2}$, leaving $\frac{1}{2}$ of hinges in the middle half.

**Not needed when hinges are not observations. (This happens for batches of sizes 4,8,12,16,20,... .)

Repeated values

When there are repetitions of the observed values, we can simplify the layout showing wings, hinges and body by giving counts of occurrences of the various values. Table 4 shows both forms for data about the number of children in each of 15 families. Notice that, in the second layout, we have used one line per value rather than one line per observation. The third panel of the same table shows the layout for what are called the "ranks" of the observations. Here the rank of the lowest observation is 1, of the next-to-lowest observation 2, and so on to the largest, here 15. The pattern of ranks shown in (c) of table 4 of course depends only on the number of observations in the batch, here 15, and not at all on what values are observed.

Comment

We have now spent considerable time on the reasons for, and notion of, a midmean and on the definition and calculation of various kinds of midmean. Doing BOTH these things is important. It is important to understand why one might want a midmean -- why it might serve us well. It is also important that, when they would serve us well, midmeans are used. As Churchill Eisenhart has succinctly put it -- in an only slightly different context:

> "The practical power of a statistical test is the product of two numbers: the mathematical power of the test and the probability that the test will be used."

In practice, neither mathematical power nor frequent use is enough alone.

table 4 of chapter 1

Three Layouts Showing Wings, Hinges and Body
for 15 Numbers of Children Per Family

	Wings	Hinges	Body	Midmeans
(a) SHOWING EACH VALUE, WITH REPETITIONS				
	5			(Easy) $\frac{20}{7} = 2.86$
	5			
	4			
		4		
			4	(Exact) $\frac{20 + \frac{1}{4}(6)}{7 + \frac{1}{4}(2)} = \frac{21.5}{7.5}$
			3	
			3	
		(keep)	3	$= 2.87$
		$\left(\frac{1}{4}\right)$	3	
			2	
			2	
		2		
	1			
	1			
	0			
Sums:	16	6	20	
Counts:	6	2	7	
(b) SHOWING COUNTS AT EACH VALUE				(Easy) $\frac{20}{7} = 2.86$
	2 at 5			
	1 at 4	1 at 4	1 at 4	(Exact) $\frac{20+1.5}{7+0.5} = 2.87$
		$(\frac{1}{4})$	4 at 3	
		1 at 2	2 at 2	
	2 at 1			
	1 at 0			
Sums:	16	6	20	
Counts:	6	2	7	
(c) SHOWING RANKS ONLY (NO INFORMATION ABOUT OBSERVATIONS)				
	13 to 15			
		12		
		$(\frac{1}{4})$	5 to 11	
		4		
	1 to 3			
Counts:	6	2	7	

(d) EXAMPLES FOR (b) ABOVE

$$20 = (1)(4) + (4)(3) + (2)(2)$$
$$6 = (1)(4) + (1)(2)$$
$$16 = (2)(5) + (1)(4) + (2)(1) + (1)(0)$$
$$1.5 = (\tfrac{1}{4})(6)$$
$$0.5 = (\tfrac{1}{4})(2)$$

1D. Clarity of residuals

There are many uses for residuals:

- for what is left of our observations after a partial description has been removed.

The kind of residuals we are now discussing, observations minus some typical value, are only the simplest instance of a very general concept that is broadly useful. Almost every type and sort of data analysis makes some use of some sort of residuals.

Table 5 sets out the residuals after fitting each of three typical values -- (i) the mean, (ii) the median, (iii) the midmean -- twice, once to our original bowling scores and once to the same scores with one score changed -- 203 replaced by 1203.

Comparing columns (3) and (4) shows us that because this extreme change of one value has violently changed the mean, every single residual from the mean has been greatly changed. If we want residuals to tell us about corresponding values, this is far from being satisfactory.

Comparing columns (5) and (6) shows us that, because this particular extreme change of one value did NOT change the median, only one residual from the median was then changed -- the one that warns us about 1203 as an observed value, the one we want to change.

Similarly, comparing columns (7) and (8) shows us that, because the midmean was affected only slightly, the residuals from the midmean were hardly changed at all, except for the one that needs to change greatly to warn us about the single wild observation.

table 5 of chapter 1

Residuals from Various Typical Values Before and After Modification of One Observation

(1) Original observations in their original order.

(2) Modified observations, one changed from 203 to 1203.

(3) and (4) Residuals from mean for unmodified and modified observations (original order).

(5) and (6) Residuals from median for unmodified and modified observations (original order).

(7) and (8) Residuals from midmean for unmodified and modified observations (original order).

(9) Unmodified observations ordered. Mean = 198 (rounded from 197.7); median = 196 (rounded from 195.5); exact midmean = 197.

(10) Modified observations ordered. Mean = 298 (rounded from 297.7); median = 196 (rounded from 195.5); exact midmean = 198 (rounded from 198.1).

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
193	193	-5	-105	-3	-3	-4	-5	184	184
214	214	+16	-84	+18	+18	+17	+16	185	185
185	185	-13	-113	-11	-11	-12	-13	191	191
198	198	0	-100	+2	+2	+1	0	192	192
203	1203	+5	+905	+7	+1007	+6	+1005	193	193
207	207	+9	-91	+11	+11	+10	+9	198	198
191	191	-7	-107	-5	-5	-6	-7	203	207
184	184	-14	-114	-12	-12	-13	-14	207	210
210	210	+12	-88	+14	+14	+13	+12	210	214
192	192	-6	-106	-4	-4	-5	-6	214	1203

Accordingly, residuals from the median or the midmean are much more likely to be helpful than residuals from the mean. For a variety of reasons which it would lead us too far astray to discuss here, the midmean appears the better choice. Accordingly, whenever the extra effort is not too large:

- we will plan to use midmeans -- or equally resistant typical values in determining our residuals.

While the idea of dividing observations into parts that can be put together with + signs -- parts for which the equation

observation = partial description + residual
= fit + residual

holds -- has been, and will continue to be, our leading idea, we do not have to be "consistent only for consistency's sweet sake" in putting this idea to work. Tools are to be used effectively.

Accordingly, it may well be reasonable to consider, and use, mismated parts of two or more different decompositions of this sort. Coat and trousers need not always match. Suppose an observation to be 43, while the mean of the batch to which it belongs is 37 and the midmean is 34, so that we have the pair of decompositions:

43 = 37 + 6 (mean + residual)

43 = 34 + 9 (midmean + residual)

There can be circumstances where we would want to use 37 as the typical value (perhaps because totals are important) and 9 as the residual (probably because we would like our residuals resistant to wild values of other observations). Such a procedure is not inconsistent, as long as we are aware of what is being done, and what may be essential for effective analysis.

Apparent consistency in exploratory data analysis is as likely to be evil as to be good. As in all exploratory fields, flexibility is usually a mark of wisdom.

1E. Comparing groups

The simplest place where we can put the idea of looking at typical values to good use is in comparisons between or among groups of observations. These comparisons can be of men with women, Mondays with Wednesdays or Thursdays, or among major league baseball clubs.

An interesting example is the annual precipitation (rain plus snow) in New York City for a set of six-year periods starting every 10 years from 1890 on. Table 6 gives the data. Looked at as a whole, it merely looks like another table of figures.

It would be helpful to look a little deeper into this array of numbers. Do there seem to be changes from decade to decade? If so, in what directions and in what amounts? Clearly it might help to look at a suitable typical value for each six-year block and see if the message -- if there be any -- is clearer.

Table 7 shows the resulting comparisons. All three of the typical values we have used are shown, each in its own

table 6 of chapter 1

Precipitation* in New York City (to nearest inch)
in Years Ending With 0,1,2,3,4 and 5

	189-	190-	191-	192-	193-	194-	195-
-0	52	42	36	49	35	45	45
-1	41	47	40	34	36	36	47
-2	39	47	38	43	39	50	46
-3	53	49	44	37	50	40	38
-4	44	42	34	38	45	52	43
-5	36	44	41	37	33	46	41

*Rain plus rain equivalent of snow.

Available from a variety of sources, including The World Almanac.

table 7 of chapter 1

Mean, Median, and Midmeans of Six-year Blocks on Rainfall Given in Table 6

	Initial Year of Block						
	1890	1900	1910	1920	1930	1940	1950
Mean	44.2	45.2	38.8	39.7	39.7	44.8	43.3
Median	42.5	45.5	39.0	37.5	37.5	45.5	44.0
Midmean	43.5	45.2	38.8	38.3	38.3	45.3	43.8

line. In this example, the mean tells the same story as the median, or the midmean: Higher typical annual rainfalls, near 44 inches, in the 1890's, 1900's, 1940's and 1950's; lower typical rainfalls, near 38 inches, in the three intervening periods.

There is now an understandable message -- 1910 to 1935 seems to have been a period of lowered rainfall. Whether or not this ought to be regarded as accident is a question we have not tried to answer -- and will not try to answer here. Our concern is that use of any one of the typical values uncovers a message in this instance -- and that their use in other instances may also uncover messages.

The first step in comparing groups will often be to calculate, and compare, typical values.

In a deeper sense, almost all typical values are for comparison. When we say that the average man in a newly discovered savage tribe is 6 feet 6 inches tall -- or 4 feet 4 inches -- the greatest impact of this figure is because the reader or listener compares it -- with other tribes, with his own country, with a mental foot rule, or with ideas of what heights are reasonable for Homo sapiens (a report of 8 feet 8 inches would arouse great doubt). There are instances of measurement for supposedly higher purposes -- for insertion into formulas expressing the fundamental laws of physics, for example. But behind the practice of such measurement is much comparison, often chains of it, as when working scales are compared with secondary standards, secondary with primary

standards, primary with national standards and so on. Typical values are usually for comparison.

1F. Some practical considerations

To have full detail about our data would sometimes be unbearable -- at other times unattainable. For 1960 in New Jersey, the Internal Revenue Service reports about rather more than 2 million 300 thousand incomes. To have detailed figures about each of them would drown us in numbers. Yet we might want to know something about New Jersey incomes.

In its published report, the Internal Revenue Service tells us about New Jersey incomes by telling us how many fall into each of 16 broad classes. Table 8 sketches the situation, indicating the ranks of the median and the hinges, and hence in which broad classes these must fall. To have 16 counts for well-chosen intervals is far better than to have 2,306,163 unsorted numbers, although we might hope to do even better with somewhat more information, say counts for 40 intervals.

What would we do, given these 16 counts, to find means, medians, midmeans and the like? Clearly we need to think things through another step or two. This example has made clear the need for thought and planning, but we will do well to use a numerically simpler example as a basis for thinking. Afterward we can return to the New Jersey incomes.

Table 9 gives hypothetical data for the age at last birthday of the members of a small college class. Here we have only 26 values, but, since the values are given in whole years -- the way most of us think about our ages -- we face the same problems as for the New Jersey incomes.

table 8 of chapter 1

Adjusted Gross Incomes for New Jersey Taxpayers in 1960

Income ($)	Number	Cumulation	Ranks	Location
None	5,873	5,873	1 to 5873	
under 1,000	217,666	223,539	5,874 to 223,539	
1,000 under 2,000	235,316	458,855	223,540 to 458,855	
2,000 under 3,000	211,042	669,897	458,856 to 669,897	**
3,000 under 4,000	239,638	909,535	669,898 to 909,535	
4,000 under 5,000	263,971	1,173,506	909,536 to 1,173,506	*
5,000 under 6,000	267,210	1,440,716	1,173,507 to 1,440,716	
6,000 under 7,000	217,673	1,658,389	1,440,717 to 1,658,389	
7,000 under 8,000	181,072	1,839,461	1,658,390 to 1,839,461	**
8,000 under 9,000	123,928	1,963,389	1,839,462 to 1,963,389	
9,000 under 10,000	76,747	3,040,136	1,963,390 to 2,040,136	
10,000 under 15,000	183,834	2,223,970	2,040,137 to 2,223,970	
15,000 under 20,000	42,159	2,266,129	2,223,971 to 2,266,129	
20,000 under 25,000	15,751	2,281,880	2,266,130 to 2,281,880	
25,000 under 50,000	19,579	2,301,459	2,281,881 to 2,301,459	
50,000 up	4,704	2,306,163	2,301,460 to 2,306,163	

*Location of median. Rank of median: $1,153,082 = \frac{1}{2}(1+2,306,163)$.

"4000 under 5000" has 243,546 = 1,153,081 - 909,535 incomes below the median, 1 at the median, and 20,424 = 1,173,506 - 1,153,082 above the median, for a total of 263,971.

**Location of hinges. $\frac{1}{4}(2,306,163) = 576,540\frac{3}{4}$, hence 576,540 observations in each wing, so that ranks of hinges are 576,541 and 1,729,623 = 2,306,163 - 576,540, respectively.

"$2000 under $3000" has 117,685 = 576,540 - 458,855 incomes in the wing, 1 in the hinge, and 93,356 = 669,897 - 576,541 in the body, for a total of 211,042.

"$7000 under $8000" has 71,233 = 1,729,623 - 1,658,390 in the body, 1 in the hinge, and 109,838 = 1,839,461 - 1,729,623 in the wing, for a total of 181,072.

Based on: Table 3, Page 28 of Statistics of Income: Supplemental Report: State and Metropolitan Area Data for Individual Tax Returns, 1959, 1960, and 1961. (Nontaxable returns of $5000 and up treated as less than $6000.)

table 9 of chapter 1

Ages of a Small College Class

"Age"*	Count	Cumul	Ranks	Interval of ages	Center age
17	1		1	17.00 to 17.99	17.5
		1			
18	11		2 to 12	18.00 to 18.99	18.5
		12			
19	7		13 to 19	19.00 to 19.99	19.5
		19			
20	4		20 to 23	20.00 to 20.99	20.5
		23			
21	2		24 to 25	21.00 to 21.99	21.5
		25			
27	1		26	27.00 to 27.99	27.5
		26			
Sum:	26				

*"Age" = Age at last birthday.

Median: Rank is $\frac{1 + 26}{2} = 13.5$, hence it falls in the 19.00 to 19.99 class.

Hinges: Since $\frac{1}{4}(26) = 6.5$, there are six values in each wing. Hence the ranks of the hinges are 7 and 20, in the 18.00 to 18.99 and 20.00 to 20.99 classes, respectively.

Note that the column of cumulative values, found by successive addition of the counts, and slipped down a half a line, makes it quite clear which interval of ranks goes with which value.

To find a detailed median, detailed hinges, or a mean, it is natural to try to spread out our values over the intervals given, since the actual ages would be spread out.

For the single student of age "27" this task is easy. Given only this information it is natural to place him at age 27.5, halfway along the "27" interval. The two students of age "21" offer a little more of a problem. Should we put them 1/3 and 2/3 of the way along the interval -- or 1/4 and 3/4?

Clearly our choice cannot make much practical difference -- moving smoothed-out values so small a fraction of an interval will not matter much. Both neatness and convenience, however, argue for 1/4 and 3/4, as we shall see.

Suppose there were -- in some other batch of data -- two ages each of "20", "21" and "22" years. If we chose 1/3 and 2/3, these would be smoothed out to

20.33, 20.67, 21.33, 21.67, 22.33, 22.67

with spacings alternating between 0.33 and 0.66. If we choose 1/4 and 3/4, they would be smoothed out to

20.25, 20.75, 21.25, 21.75, 22.25, 22.75

a nice even sequence. Neatness favors 1/4 and 3/4.

It would be nice to have these smoothing-out fractions taken care of almost automatically, saving as much thought as possible. This we can do, if we have chosen 1/4 and 3/4, by using ranks. Table **10** shows the results, for part of table 9, of assigning ranks not only to observations but also to the cutting values that separate one "age" from another. Once cutting values have ranks, we can use a simple uniformly changing relationship of ranks to values each "age". For "21" for instance, the cutting values of 21 and 22 are ranked 23.5 and 25.5 so that we move 0.5 units of value for each unit of rank. The actual ranks of 24 and 25 thus correspond to values of 21.25 and 21.75, as we would wish.

Having learned how to smooth out values, we can turn to finding medians, hinges, and midmeans. Table **11** shows the wings, hinges, and body layouts for (a) the ranks of the 26 "ages", (b) the values of the 26 "ages" and (c) the smoothed-out ages. Wherever all instances of an "age" fall in a wing, or a hinge, or the body, we assign a midvalue. Otherwise we think about the smoothed-out values in more detail. In assigning a single representative value to the five "ages" of 18 that fall in the lower wing and which would have received individual smoothed-out values of

18.05, 18.14, 18.23, 18.32, 18.41

we have taken

$$18.23 = \frac{1}{2}(18.05 + 18.41)$$

$$= \frac{1}{5}(18.05 + 18.14 + 18.23 + 18.32 + 18.41)$$

table 10 of chapter 1

Use of Ranks for Cutting Points to Assign Smoothed-out Values
(Applied to Part of the Data of table 9)

Cutting Point	Range of ages	Rank	Smoothed-out value
19		12.5	
	19.00 to 19.99	13	19.07
	"	14	19.21
	"	15	19.35
	"	16	19.50
	"	17	19.65
	"	18	19.79
	"	19	19.93
20		19.5	
	20.00 to 20.99	20	20.12
	"	21	20.38
	"	22	20.62
	"	23	20.88
21		23.5	
	21.00 to 21.99	24	21.25
	"	25	21.75
22		25.5	

Examples: $19.35 = 19 + \frac{15 - 12.5}{19.5 - 12.5}(20-19)$

$$20.62 = 20 + \left(\frac{22 - 19.5}{23.5 - 19.5}\right)(21-20)$$

table 11 of chapter 1

Layouts for 26 Observations
(Making Use of tables 9 and 10)

	Wings	Hinges	Body	Ages
(a) LAYOUT OF RANKS FOR DATA IN TABLE 9				
1.5	1			"17"
12.5	2 to 6	7	8 to 12	"18"
19.5			13 to 19	"19"
23.5	21 to 23	20		"20"
25.5	24 to 25			"21"
	26			"27"
(b) LAYOUT OF "AGES" FOR DATA IN TABLE 9				
1.5	1 at "17"			
12.5	5 at "18"	1 at "18"	5 at "18"	
19.5			7 at "19"	
23.5	3 at "20"	1 at "20"		
25.5	2 at "21"			
	1 at "27"			
(c) LAYOUT WITH SMOOTHED-OUT AGES				
1.5	1 at 17.50			
12.5	5 at 18.23	1 at 18.50	5 at 18.77	
19.5			7 at 19.50	
23.5	3 at 20.63	1 at 20.12		
25.5	2 at 21.50			
	1 at 27.50			

(d) EXAMPLES

$$18.50 = 18 + \frac{7 - 1.5}{12.5 - 1.5}(19-18)$$

$$20.12 = 20 + \frac{20 - 19.5}{23.5 - 19.5}(21-20)$$

$$18.23 = \tfrac{1}{2}(18.05 + 18.41), \text{ where}$$

$$18.05 = 18 + \frac{2 - 1.5}{12.5 - 1.5}(19-18)$$

$$18.41 = 18 + \frac{6 - 1.5}{12.5 - 1.5}(19-18)$$

$$20.63 = \tfrac{1}{2}(20.38+20.88)$$

table 11 (Cont'd) - 2

(e) MEDIAN AND SKELETON MIDMEAN

Median falls at rank 13.5, its smoothed-out value is

$$19 + \frac{13.5 - 12.5}{19.5 - 12.5}(20-19) = 19.14$$

$$\text{Skeleton midmean} = \frac{18.50 + 19.14 + 19.14 + 20.12}{4}$$

$$= 19.22$$

as the common value. (The two expressions are equal because the values are equally spaced.)

Note that we would have found almost the same representative values if we had said that these five values were between 18.00 and 18.50, and taken 18.25. (If we had tried this for the three values that gave 20.63, we would have found 20.56, a somewhat larger discrepancy.) Once there are more than a few observations per class, this simpler calculation gives adequate accuracy.

If we had needed only the skeleton midmean we would have calculated only 18.50, 19.14, and 20.12 in order to find 19.22. Only a quite limited amount of arithmetic is needed to do this.

Table 12 repeats the layout for the smoothed-out values from table 11(c) and completes the calculation of midmeans and mean.

1G. Back to incomes

We are now ready to return to our 2,306,163 New Jersey incomes. Table 13 gives the wings, hinges, and body display for this data. Checking the calculations will require repeated reference to table 8, but will show that up to this point, while the numbers are larger, no more arithmetic operations need be made than in our small example of 26 ages.

The assignment of midvalues by uniform spreading deserves some attention, both because of obvious difficulties and because the Internal Revenue Service provides data sufficient to check ourselves.

2I

table 12 of chapter 1

The Layout for Smoothed-out Values Repeated and the Calculation of Midmeans and Mean

(a) LAYOUT OF 26 SMOOTHED-OUT VALUES, REPEATED FROM TABLE 11(c)

Wings	Hinges	Body
1 at 17.50		
5 at 18.23	1 at 18.50	5 at 18.77
		7 at 19.50
3 at 20.63	1 at 20.12	
2 at 21.50		
1 at 27.50		
Sums: 241.04	38.62	230.35
ounts: 12	2	12

Midmeans and mean

Easy midmean:

$$\frac{230.35}{12} = 19.2$$

Exact midmean:

$$\frac{203.35 + 19.31}{12 + 1} = 19.2$$

Mean:

$$\frac{230.35 + 38.62 + 241.04}{12 + 2 + 12} = 19.6$$

(b) EXAMPLES

$$230.35 = (5)(18.77) + (7)(19.50)$$

$$38.62 = 18.50 + 20.12$$

$$241.04 = 17.50 + (5)(18.23) + (3)(20.63) + (2)(21.50) + 27.50$$

table 13 of chapter 1

Wings, Hinges, Body Display of Numbers of Incomes
for the Data of table 8

Income ($)	Wings		Hinges	Body
None	5,873 at	0		
under 1,000	217,666 at	500		
1,000 under 2,000	235,316 at	1,500		
2,000 under 3,000	117,685 at	2,278	1 at $2557	93,356 at $2,778
3,000 under 4,000				239,638 at $3,500
4,000 under 5,000				263,971 at $4,500
5,000 under 6,000				267,210 at $5,500
6,000 under 7,000				217,673 at $6,500
7,000 under 8,000	109,838 at	$7,697	1 at $7393	71,233 at $7,197
8,000 under 9,000	123,928 at	$8,500		
9,000 under 10,000	76,747 at	$9,500		
10,000 under 15,000	183,834 at	$12,500		
15,000 under 20,000	42,159 at	$17,500		
20,000 under 25,000	15,751 at	$22,500		
25,000 under 50,000	19,579 at	$37,500		
50,000 up	4,704 at	$ what?		
Sums:	???		$9,949	$5,683,138,869
Counts:	1,153,080		2	1,153,081

Examples: $$\$2,557 = \$2,000 + \frac{576,541 - 458,855.5}{669,897.5 - 458,855.5}(\$3000 - \$2000)$$

$$\$7,393 = \$7,000 + \frac{1,729,623 - 1,658,389.5}{1,839,461.5 - 1,658,389.5}(\$8,000-\$7,000)$$

$$\$2,278 = \tfrac{1}{2}(\$2,000 + \$2,557)$$

$$\$2,778 = \tfrac{1}{2}(\$2,557 + \$3,000)$$

Median:

$$\$4,923 = \$4,000 + \frac{1,153,082 - 909,535.5}{1,173,506.5 - 909,535.5}(\$5000-\$4000)$$

Skeleton midmean:

$$\$4,949 = \tfrac{1}{4}(\$2,557 + \$4,923 + \$4,923 + \$7,393)$$

Looking only at table 13, we are entitled to wonder:

- whether 5,873 incomes were exactly zero.
- whether the 217,666 incomes under $1000 are nearly uniformly spread from $0 to $1000.
- whether the 19,579 incomes between $25,000 and $50,000 are nearly uniformly spread.
- what to do with the open-ended class "$50,000 up".

As is so often the case, none of these most obvious questions affect the calculation of any of the midmeans. The skeleton midmean is calculated in the table. The easy midmean requires the sums shown in the table, one of which

$$(93,356)(2778) + (239,638)(3500) + (263,971)(4500) + (267,210)(5500) + (217,673)(6500) + (71,233)(7197) = 5,683,138,869$$

can be done much easier if we can use some sort of desk calculator.

We thus find that the easy midmean of the smoothed-out values is

$$\$4929 = \frac{5,683,138,869}{1,153,081}$$

while the exact midmean must differ by so little that we will not bother to work it out. The agreement with the median is closer than with the skeleton midmean, mainly because the classes below $4000 are substantially larger than those above $6000.

The actual means for the 16 income classes, as given by the Internal Revenue Service, are set out in table 14 and compared with those assigned. We see that for the 10 income classes that are $1000 wide the actual means agree with the smoothed out means to ±$29. This seems surprisingly good agreement. As we go to intervals above $10,000, we see that the actual means fall well below the smoothed-out means, as we would expect. In general, the smoothing-out process works quite well, even with this data.

Stretching a point

The technique of smoothing-out reported values and then calculating medians or midmeans of the stretched-out values is clearly very useful where it is necessary, as for the New Jersey incomes. It may or may not be necessary, for the student's ages, but it is surely both appropriate and helpful.

More surprising, however, are the cases where it is helpful, though not strictly appropriate. The County and City Data Book, published by the Bureau of the Census every few years, gives much information about every county in the United States, including the "median number of years of school completed" of which the 1967 book says at page xxi:

table 14 of chapter 1

Mean Incomes for the Various Classes

Class	Count	Income			
		Total	Mean	(*)	diff
None	5,873	-	-	0	
under $1,000	217,666	113,822	523	500	+23
1,000 under 2,000	235,316	346,213	1,471	1500	-29
2,000 under 3,000	211,042	529,397	2,508	2500**	+8
3,000 under 4,000	239,638	840,063	3,506	3500	+6
4,000 under 5,000	263,971	1,191,632	4,514	4500	+14
5,000 under 6,000	267,210	1,477,278	5,529	5500	+29
6,000 under 7,000	217,673	1,411,283	6,484	6500	-16
7,000 under 8,000	181,072	1,355,260	7,485	7500**	-15
8,000 under 9,000	123,928	1,050,723	8,478	8500	-22
9,000 under 10,000	76,747	727,966	9,485	9500	-15
10,000 under 15,000	183,834	2,170,372	11,806	12,500	-694
15,000 under 20,000	42,159	720,375	17,087	17,500	-413
20,000 under 25,000	15,751	347,596	22,068	22,500	-432
25,000 under 50,000	19,579	637,352	32,553	37,500	-4947
$50,000 up	4,704	398,820	84,783	what?	

*Assigned in table 13.

**Split in pieces in table 13, this is value that would have been assigned.

"The median number of school years completed is the value which divides the population group into two equal parts -- one half having completed more, and the other half less, schooling than the median".

Presumably, the number of school years completed comes in integer years, either for all by definition or for great majority by practice. None of us believes, I am sure, that the Census collects -- or should collect -- data about school years completed in fractions of a year. Yet the resulting medians are given in years and tenths. Clearly the medians reported must be medians for some sort of smoothed-out values.

As a measure of more-or-less typical amounts of schooling, the median of the smoothed-out values is more responsive to small changes from one county to another than the median of the raw values would be. While a purist might try to object, the practical consensus will be that spreading out these values helps.

This can be true in more extreme situations, like the numbers of children per family set out in table 4, earlier, where the median of the smoothed-out numbers of children can be found from

rank	value
6.5	2.5
10.5	3.5

to be 2.88.

So far as the formal definition of a median goes, it has to be an actual value. No family admits to having 2.88 children. Yet smoothed-out medians are often more useful than actual-value medians. The comparison of two large groups of families implied by smoothed-out medians of 2.6 and 2.9 children, respectively, is much more useful than the less informative actual medians of 3 and 3.

1H. How far have we come?

In this chapter we have dealt -- more or less thoroughly, as the case may be -- with nine important ideas:

- typical values (of three kinds)
- partial descriptions (of one kind)
- wild values (a regrettable commonplace)
- resistance (to confusion; one case)
- residuals (of one sort -- deviation from a typical value)
- clarity (of representation; one case)
- comparison (a central purpose)
- penetrating examination (our theme).

Let us then try to summarize where we stand with respect to each.

Given a batch of numbers, a typical value calculated from them can serve as a summary for varied purposes, including:

- as an indication of level
- as a partial description generating residuals.

Any such value is clearly a partial description of all the values in the batch.

We have defined and calculated examples of: the arithmetic mean, the median, and various kinds of midmean (easy, exact, and skeleton). Each has its advantages, only a few of which we have mentioned.

We recognize the all too frequent occurrence of "wild values" . We recognize that there are many reasons for this, so that it is almost always important to try to understand the "why" of a wild value, rather than merely labeling as such. Is it an accident, is it merely erratic, or is it a clue to some unanticipated phenomenon? Is it "way out" or only "far out"? We recognize that these questions cannot always be answered.

ONE good property of some typical values is their resistance to the influence of one or a few wild values. Just as for any way of providing a partial description, resistance is a valuable property for a way of defining a typical value. Here median and midmean are much better than arithmetic mean.

If we are to have as broadly applicable a standard procedure as possible, then, we should use a midmean as a satisfactory compromise. Which midmean we use is of much less importance. Ease of hand computation often indicates using the easy midmean for small batches, but the skeleton midmean for large ones. An "automatic" computer might as well use the exact midmean.

Given any partial description of the data, we can always try to look deeper by forming the differences between what was observed and our partial description. In practice, we usually do this by subtracting one from the other -- and calling the difference a residual . (There remains a question of how we express both before we subtract.)

2I

So far, the only partial descriptions we have met are typical values. Accordingly, the only residuals we have yet met are of the form:

(observed value) - (typical value of batch)

We shall meet many others.

Once we form residuals, whose purpose is always to let us look deeper into the structure (real or apparent) of the data, we must try to avoid misleading appearances. If each residual could only tell us about the corresponding observed value, all would be well. But our partial description must come from somewhere, almost always from data. If we are to hold down confusion, if there is to be clarity in the connection of residual to observation, our partial description needs to be calculated in a resistant way. In particular, we will want to use resistant typical values, like the median and the midmean, in forming residuals.

The basic use of summarization is comparison . Sometimes it is comparison with an anticipated or theoretical value, sometimes it is comparison among groups. (To say that a college freshmen class averages 17 years and 6 months old is an easy example to think about.) Measurement is always for some sort of comparison.

Even in this first chapter, our stress has been where it will steadily be: on looking deeper, on more penetrating examination of our data. We would not stop with calculating typical values. Rather we pressed on to finding residuals and looking for stragglers. This is the pattern we shall follow; this is what we most need to learn to do.

*****

1K. <u>How to put in order</u>

If any of us is given three numbers to put in order, the task is easy. Ten numbers strain -- at least to its utmost -- the technique of looking up and down the column for the next value. There ought to be a simple way to order numbers -- and there is one.

Let us look at it, beginning with the 10 bowling scores. A quick look shows that they begin with 18, 19, 20 or 21. So we write out the following pattern:

18	
19	
20	
21	

where we shall mark down each value, as we come to it, by putting its last digit after the first digits already in place. Since 193 and 214 were the first two given, we find first

18	
19	3
20	
21	

and then

18	
19	3
20	
21	4

It may help our thinking to call 19 and 21 "stems" and 3 and 4 "leaves". We will call such a layout a "stem-and-leaf" array.

Once the process is completed we have this stem-and-leaf array

18	54
19	3812
20	37
21	40

and have now only to arrange the single digits in each row. (Scanning through a row for single digits is usually easy; if there are too many to count easily, we can turn to Section 2K and use its technique.)

If we wish to keep everything perfectly easy to check, we can write down the little table again, with the last digits in order. In our example we get:

18	45
19	1238
20	37
21	04

from which it is trivial to write out 184,185,191,192,193,198, 203,207,210,214.

In some instances we may want to let pairs of last digits play the role that single digits play in the simple technique. Commas are then needed between entries.

One of the most common situations where this arises is when numbers are usually small, often moderate, and occasionally large. Table 15 shows how the procedure operates in such a case.

table 15 of chapter 1

Collection of Mixed Values, Occasionally Large, As A Preliminary to Ordering (or Grouping)

(a) The numbers: 3,17,26,19,47,69,23,2,5,68,93,27,3,7, 4,11,87,42,157,23,49,193,24,35,243,512,47,25,26,14,8,4,3,12, 11,24,29,81,35,5,7.

(b) The collection:

0	32537484357
1	791421
2	637345649
3	55
4	7297
5	
6	98
7	
8	71
9	3
1	57,93,
2	43,
3	
4	
5	10,

Note that two vertical rules, as in the last five lines, indicate two more digits to follow.

Up or down?

This process introduces us to a problem that seems to have no simple answer -- a problem that arises with many layouts, including those that show wings, hinges, and body. Shall the numbers "begin" at the top of the layout and increase downward, as we number lines on a page? Or shall they "begin" at the bottom and increase upward, as we plot numbers?

The best the writer seems able to do is to compromise. He proposes:

- when the numbers do not cross zero, to go down the page from small to large.

- when both negative and positive values appear, to go up the page from minus to plus.

The reader is encouraged to experiment.

In any case, we want to remember that:

- in a stem-and-leaf layout, there are separate lines for +0 and -0.

- a firm horizontal line is needed where we cross zero (or, as in table 15 change the mode of entry).

Other uses

Clearly the stem-and-leaf collection procedure can be used for other purposes than ordering. Some examples are:

- if we want to count numbers of occurrences in certain ranges -- especially when the ranges cannot or should not be fixed until we understand the pattern of variation of the numbers.

- especially when the same number of right-hand digits are stacked up throughout, the collection gives a useful visual impression of the numbers' pattern of variation. (Even the impression given in the situation of table 15, where the space taken up by a single original number is not constant, is helpful.)

Like any hand procedure, this ordering procedure is likely to need checking. Even though stem-and-leaf arraying leads to fewer errors than do most hand processes, there are going to be cases where the count of the number of actual entries in our stem-and-leaf array will not be the same as in the original data, showing that we have slipped, and must find the error.

Doing the whole array over again is unnecessary, and may not help. We can do better, in an easier way. We merely have to go through the original data again in the same order, placing a dot over each digit (or multidigit entry) in the original stem-and-leaf array as we come to it. So long as there has been no error, the digit to be dotted will be the first still-undotted digit in its line. Errors are easily detected, and we can usually even tell how they happened.

All in all, this simple collection process is a very great convenience.

1M. Typical values for weighted numbers

Not infrequently we wish to give more attention to certain observations in a batch and less to others. This may be because some observations summarize more extensive bodies of data, others less. Or because some observations involve better measurement techniques, others poorer. Or for any of many other reasons.

The simplest principle in such situations is that of simulated repetition. This means treating certain observations as if there were two or more observations whose numerical expression was the same.

Consider a very simple situation, where the numerical observations are 7, 9, 3 and 6, and where we desire to weight 9 doubly and 6 trebly. We have only to act as if the observations were 7, 9, 9, 3, 6, 6 and 6, so far as typical values go. Table 16 shows the calculation of mean, median, and midmean. The first two panels are exactly the same, emphasizing the fact that smoothing out a weighted value to different numbers makes no sense when these numbers are duplicated only to reflect weights.

So long as we are not going to smooth out our values, there is no difficulty in following the same sort of calculation with weights that are not whole numbers. The computations stay just as easy.

table 16 of chapter 1

Layout for Midmean of Four Weighted Observations

	Wings	Hinges	Body	Midmeans
(a)	RAW VALUES			Easy midmean
	1 at 9	1 at 9		$\frac{19}{3} = 6.3$
			1 at 7	
		(keep $\frac{1}{4}$) 1 at 6	2 at 6	Exact midmean
	1 at 3			$\frac{19 + 3.75}{3 + 0.5} = 6.5$
Sums:	12	15	19	
Counts:	2	2	3	
(b)	SMOOTHED-OUT VALUES			Easy midmean
	1 at 9	1 at 9		$\frac{19}{3} = 6.3$
		(keep $\frac{1}{4}$)	1 at 7	
			2 at 6	Exact midmean
		1 at 6		$\frac{19 + 3.75}{3 + 0.5} = 6.5$
	1 at 3			
Sums:	12	15	19	
Counts:	2	2	3	

(c) OTHER TYPICALS

Median = 6

$$\text{Mean} = \frac{12+15+19}{2+2+3} = 6.57$$

*****

1P. <u>IEDA processor</u>

*****

1R. Arithmetic of splitting

In preparation for what is to come later, let us consider the splitting-up of a simple set of data into pieces that can be reunited by plus signs to give what we started with. Consider the 10 bowling scores, 193, 214, 185, 198, 203, 207, 191, 184, 210 and 192. Their mean was 198 (rounded from 197.7), so one such split-up goes as follows:

$$193 = (198) + (-5)$$

$$214 = (198) + (+16)$$

$$185 = (198) + (-13)$$

$$198 = (198) + (0)$$

$$203 = (198) + (+5)$$

$$207 = (198) + (+9)$$

$$191 = (198) + (-7)$$

$$184 = (198) + (-14)$$

$$210 = (198) + (+12)$$

$$192 = (198) + (-6)$$

The pieces "left over" once the mean is extracted are called residuals rather than "leftovers", presumably because Latin stems are more scientific (and, perhaps, because they offer a greater chance of generating similar-looking words in various languages).

We can split out the median -- or the midmean -- in just the same way. For the same set of scores, splitting out the midmean gives:

$$193 = (197.3) + (-4.3)$$
$$214 = (197.3) + (+16.7)$$
$$185 = (197.3) + (-12.3)$$
$$198 = (197.3) + (+0.7)$$
$$203 = (197.3) + (+5.7)$$
$$207 = (197.3) + (+9.7)$$
$$191 = (197.3) + (-6.3)$$
$$184 = (197.3) + (-13.3)$$
$$210 = (197.3) + (+12.7)$$
$$192 = (197.3) + (-5.3)$$

The right-hand column here is just as much "left over" as the right-hand column above. These pieces, too, are properly called residuals.

In general, residuals are what is left over after one has subtracted the results of some systematic process of partial description. (The use of judgment can be, for us at least, part of a systematic process of description. If a man looks at the 10 bowling scores, says to himself "they run about 200" and starts doing 193 = (200) + (-7), 214 = (200) + (+14), 185 = (200) + (-15), we are glad to call -7, +14, -15, etc. residuals.)

As we go on, we will find occasion for more complex breakdowns into pieces that can be put back together with + signs.

1W. Background references

Some discussion of the mean will be found in almost any textbook of statistics. Most of these discussions will help you little with the point of view we are taking here. (Once you get this point of view clearly in mind, reading some of these discussions with a view to how you might want to change them could be profitable.)

Except that there will be fewer places where it is discussed, the same can be said for the median.

You will find almost nothing about the midmean in print. (((Statisticians who have strayed into reading this may be able to extract some nourishment from papers (i) by Crow and Siddiqui, at pages 353-389, and (ii) by Siddiqui and Raghunandanan at pages 950 to 953, both of volume 62 (1967) of the Journal of the American Statistical Association. They will have to read carefully and reword conclusions. They need to be warned that the aspect with which these writers are concerned is important, but is not our main issue.)))

*****

1Y. Problems

The problems provided here are divided into three categories:

Problems for hand arithmetic

Note: The annual rainfalls and winter snowfalls in New York City (problems 1-4,7-8) can be found in various sources, including The World Almanac.

(1) The annual rainfalls in New York City for the six years 1960 to 1965 were 46, 39, 37, 34, 33, and 26 inches (in that order, to the nearest inch). Compute the mean, median and midmean, list together with the values from table 7, and comment on the result.

(2) Discover the annual rainfall for New York City for the six-year periods 1895-1900, 1905-1910, 1915-1920, 1925-1930, 1935-1940, 1945-1950, 1955-1960. Choose a typical value and calculate its value for each batch. Compare with the results of table 7. Comment.

(3) The snowfalls in New York City for 20 consecutive winters from 1918-19 to 1937-38 were (in inches) 3.5, 55.4, 18.2, 29.7, 55.2, 26.3, 27.9, 35.8, 21.9, 14.3, 13.3, 13.5, 9.7, 5.1, 24.5, 53.1, 29.0, 32.8, 11.9, 13.9. Find the skeleton midmean. Do you think other typical values would be very different?

(4) The snowfalls in New York City for 20 consecutive winters from 1938-39 to 1957-58 were (in inches) 31.9, 22.2, 35.0, 10.2, 27.6, 26.0, 26.7, 26.6, 33.2, 61.5, 43.0, 10.4, 10.9, 14.4, 9.1, 17.1, 10.9, 29.8, 19.1, 37.9. Find the skeleton midmean and compare with the result of problem 3.

(5) Table 17 gives the % of major party vote for Nixon and Goldwater in eight southwestern states in 1960 and 1964. Find the differences "Nixon-Goldwater" and then the mean, median, and midmean of these differences.

(6) During the winter of 1893-94, Lord Rayleigh was investigating the density of nitrogen from various sources. He had previously found indications of a discrepancy between the densities of nitrogen produced by removing the oxygen from air and nitrogen produced by decomposition of a chemical compound. The 1893-94 results established this difference with great definiteness and led him to investigate further the composition of air chemically freed of oxygen. This led to the discovery of argon, a new gaseous element.

Rayleigh's observations are summarized in table 18. Calculate the midmean weight for "nitrogen" from air. Calculate the midmean weight for "nitrogen" prepared from NO, N_2O, or NH_4NO_2. Compare and comment.

(7) In a study of cost of plastering in 1923 by the Bureau of Labor Statistics, plastering costs in dollars per square yard were obtained for 19 jobs in Atlanta, Georgia. These were 0.45, 0.16, 0.16, 0.15, 0.14, 0.15, 0.12, 0.29, 0.31, 0.20, 0.30, 0.27, 0.28, 0.40, 0.35, 0.14, 0.14, 0.18, 0.19. Calculate the skeleton midmean of these 19 values.

(8) Goulden gives the % of carotene in the flour and in the grain for 139 varieties of wheat. The percent carotene in the flour for the first 99 varieties was 2.39,3.11,2.15,1.96,

table 17 of chapter 1

% of Major Party Vote for Republican Presidential Candidates in Eight Southwestern States

	% for Nixon in 1960	% for Goldwater in 1964
Arizona	55.6%	50.5%
California	50.3%	40.8%
Colorado	54.9%	38.4%
Nevada	48.8%	36.1%
New Mexico	49.6%	40.6%
Oklahoma	59.0%	44.3%
Texas	50.5%	36.6%
Utah	54.8%	45.3%
(All U.S.)	(49.9%)	(38.7%)

From Richard M. Scammon 1965. America at the Polls: A Handbook of Presidential Election Statistics 1920-1964. Pittsburgh: University of Pittsburgh Press: 521 pp.

table 18 of chapter 1

Rayleigh's Weights of a Standard Volume of "Nitrogen"

Date	Source of "nitrogen"		Weight
	Origin	Purifying agent	
29 Nov. '93	NO	hot iron	2.30143
5 Dec. '93	"	"	2.29816
6 Dec. '93	"	"	2.30182
8 Dec. '93	"	"	2.29890
12 Dec. '93	Air		2.31017
14 Dec. '93	"	"	2.30986
19 Dec. '93	"	"	2.31010
22 Dec. '93	"	"	2.31001
26 Dec. '93	N_2O	"	2.29889
28 Dec. '93	"	"	2.29940
9 Jan. '94	NH_4NO_2	"	2.29849
13 Jan. '94	"	"	2.29889
27 Jan. '94	Air	Ferrous hydrate	2.31024
30 Jan. '94	"	"	2.31030
1 Feb. '94	"	"	2.31028

From Lord Rayleigh 1894. On an anomaly encountered in determinations of the density of nitrogen gas. Proc. Roy. Soc. (Lond.) 55, 340-344 (also in his Scientific Papers Vol. 4, pp. 104-108).

2.02,1.76,2.10,2.12,2.28,1.86,2.60,2.11,2.30,1.80,2.00,2.05, 2.09,2.33,2.29,2.30,1.97,2.36,1.73,1.72,1.70,1.63,1.93,1.50, 1.77,1.60,2.31,2.17,2.10,2.90,2.17,2.15,2.01,2.35,2.35,2.00, 2.18,2.47,2.25,1.77,1.68,2.46,1.86,1.71,1.93,1.81,1.89,1.65, 1.93,2.12,2.25,1.92,2.25,2.25,1.65,1.63,1.70,1.61,1.83,1.60, 1.37,1.96,1.88,1.92,1.89,1.99,1.82,2.12,2.16,2.14,1.63,2.76, 2.07,1.67,2.78,3.40,3.67,2.41,2.23,3.07,2.22,2.55,2.12,1.94, 1.95,1.59,2.00,1.77,1.98,1.97,1.97,1.83,2.00,1.96,2.00. Find the skeleton midmean.

(From page 65 of C. H. Goulden 1939. Methods of Statistical Analysis. New York: John Wiley and Sons: 277 pp.)

Problems for automatic computation

To be issued separately.

Theoretical problems

(21) In a batch of eight values, what are the largest and smallest possible numbers of observations that can be greater than the mean? -than the median? -than the exact midmean? -than the skeleton midmean? Can you prove your statements?

(22) Can you generalize the results of problem 21? How far?

(23) What light do you think (i) the conclusions and (ii) the detailed workings of problems 21 and 22 cast on behavior of means, medians, and midmeans in the presence of wild values?

(24) If the four values generating table 16 -- namely 7, 9, 3 and 6 -- are to be weighted in all possible ways, what values for the median are possible? What values for the midmean? What values for the mean? What sets of weights prove your statements?